Language Technologies Institute

Carnegie Mellon University

Advanced Topics in Multimodal Machine Learning (11-877)

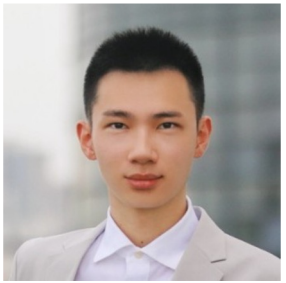
Lecture 1: Introduction

Paul Liang and Daniel Fried

Spring 2024

Your Teaching Team This Semester (11-877, Spring 2024)

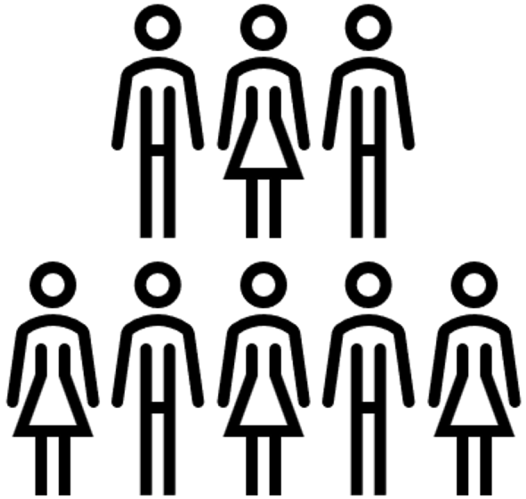
Daniel Fried
dfried@cs.cmu.edu
Course instructor



Haofei Yu
haofeiy@andrew.cmu.edu
Teaching Assistant

Paul Liang
pliang@cs.cmu.edu
Course instructor

Time for Introductions!



Your name, department and programs

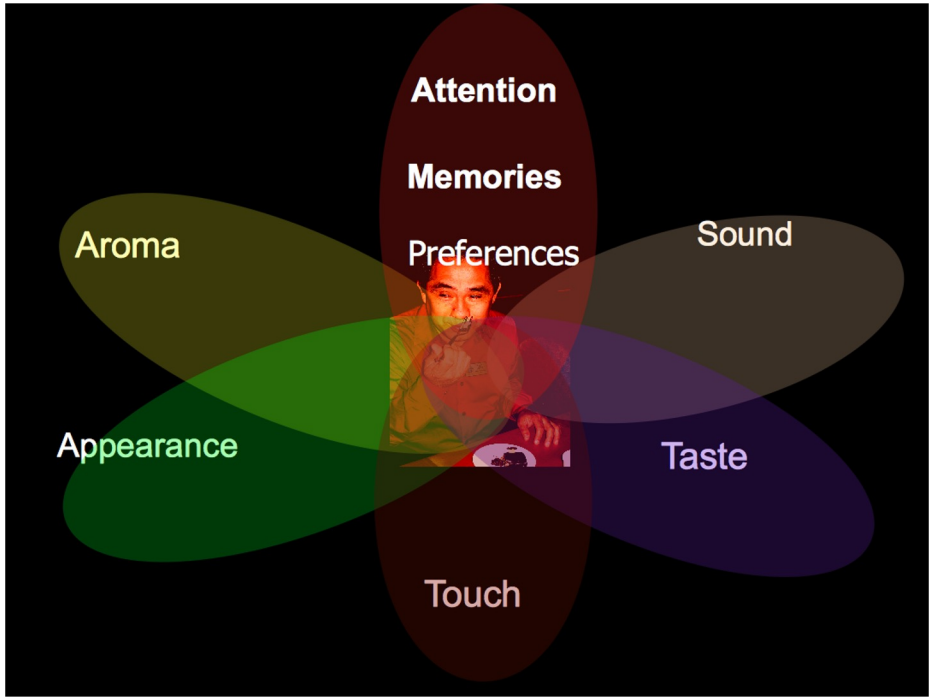
Your favorite modality(ies)!

Previous research experience in multimodal

Why are you interested in this course?

What is Multimodal?

What is Multimodal?



Sensory Modalities

Multimodal Behaviors and Signals

Language

- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

Acoustic

- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual

- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Touch

- **Haptics**
- **Motion**

Physiological

- **Skin conductance**
- **Electrocardiogram**

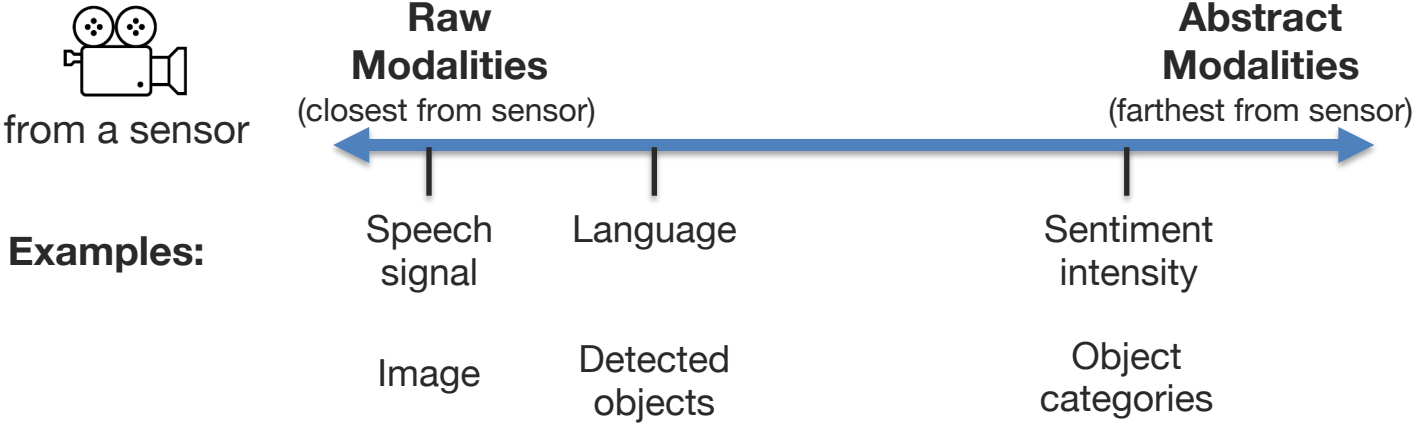
Mobile

- **GPS location**
- **Accelerometer**
- **Light sensors**

What is a Modality?

Modality

Modality refers to the way in which something expressed or perceived.



What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

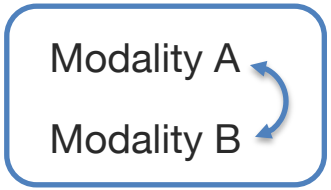
A research-oriented definition…

***Multimodal* is the science of**

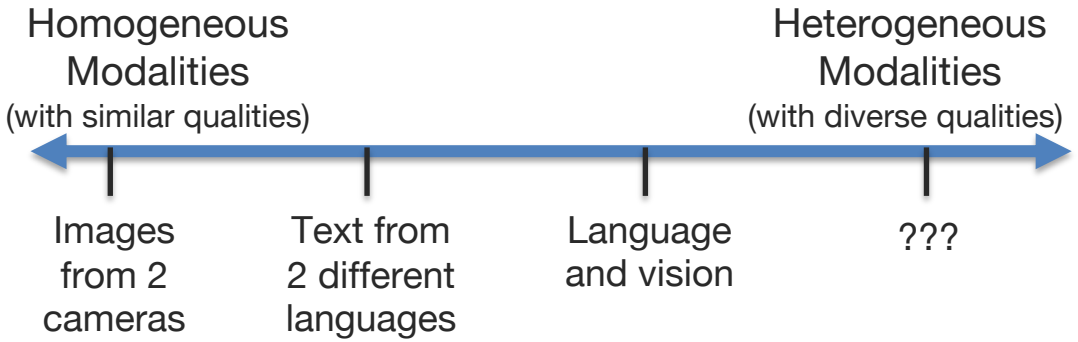
heterogeneous and interconnected data

Principle 1: Heterogeneous Modalities

Information present in different modalities will often show diverse qualities, structures and representations.



Examples:



Abstract modalities are more likely to be homogeneous

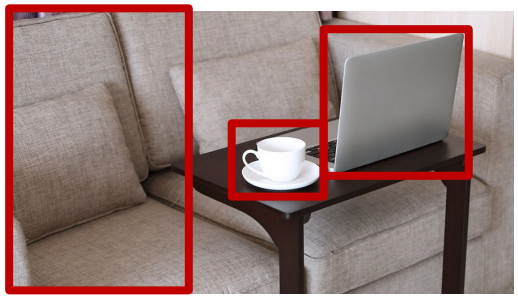
Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.



*A **teacup** on the **right** of a **laptop** in a **clean room**.*

1. **Element representations:** discrete, continuous, granularity

{teacup, right, laptop, clean, room}

Dimensions of Heterogeneity

Modality A ⟺ Modality B

1. **Element representations:**
   Discrete, continuous, granularity

2. **Element distributions:**
   Density, frequency

3. **Structure:**
   Temporal, spatial, latent, explicit

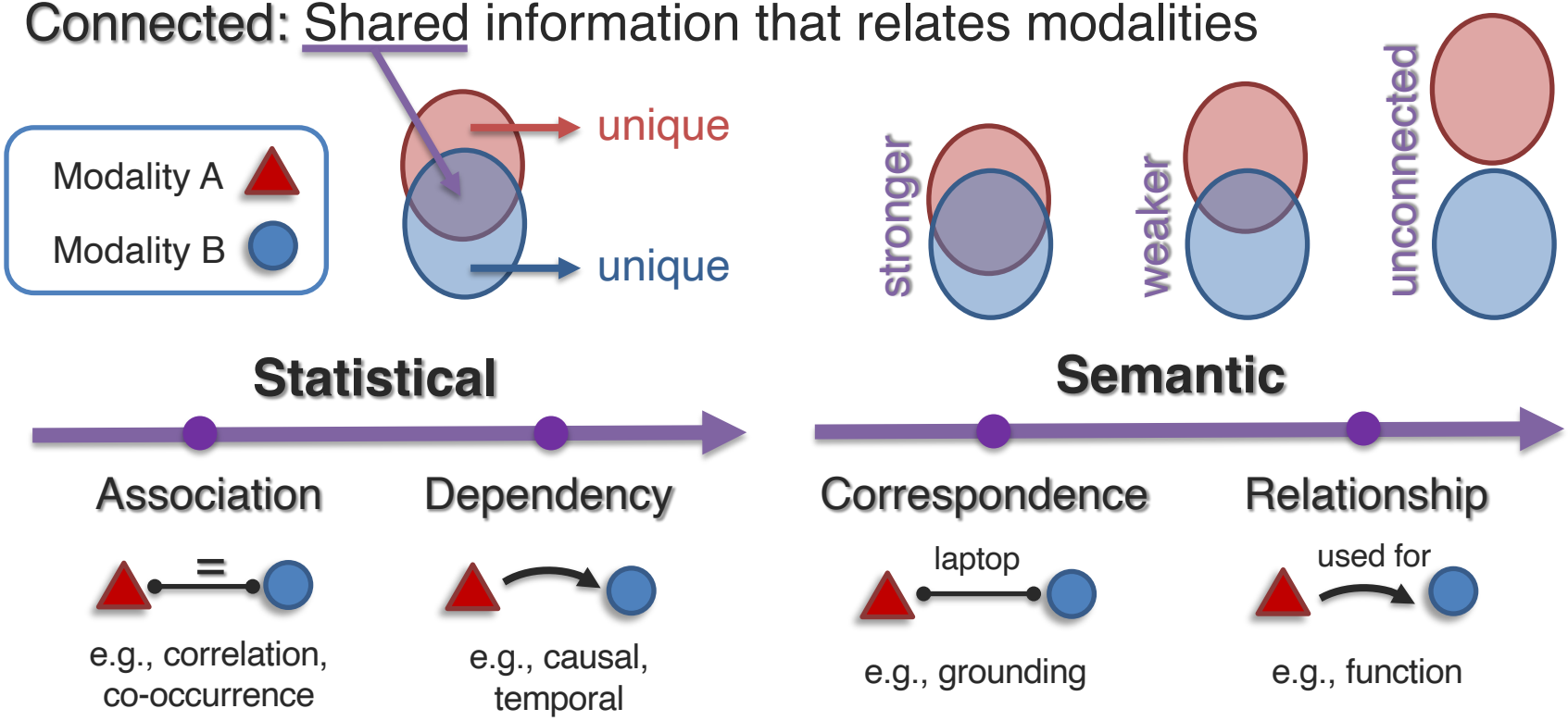
4. **Information:**
   Abstraction, entropy

5. **Noise:**
   Uncertainty, noise, missing data

6. **Relevance:**
   Task, context dependence

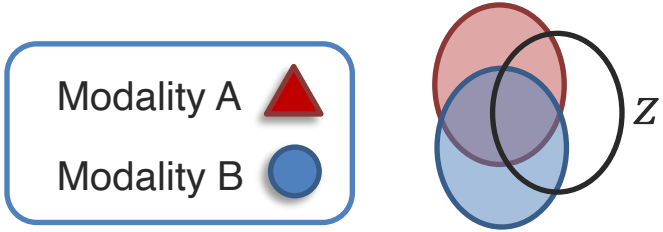
Principle 2: Modalities are Connected

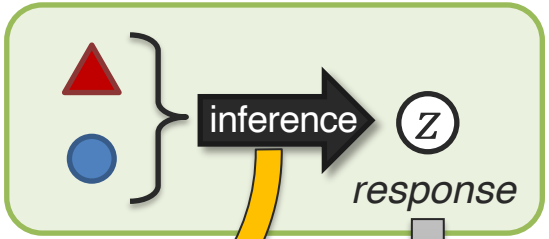
Connected: Shared information that relates modalities



Principle 3: Modalities Interact

Interacting: process affecting each modality, creating new response





Interactions happen during inference!

"Inference" examples:

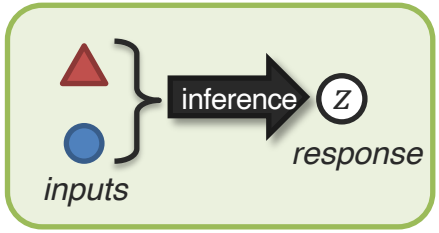
- Representation fusion
- Prediction task
- Modality translation

representation

$\hat{y}$ prediction

modality C

Taxonomy of Interaction Responses – A Behavioral Science View



Multimodal Communication

	signal	response	signal	response	
Redundancy	a	□	a+b	□	Equivalence
	b	□	a+b	▭	Enhancement
Nonredundancy	a	□	a+b	□ and ○	Independence
	b	○	a+b	□	Dominance
			a+b	▭ (or □)	Modulation
			a+b	△	Emergence

Partan and Marler (2005). *Issues in the classification of multimodal communication signals*. American Naturalist, 166(2)

Multimodal Technical Challenges – Surveys, Tutorials and Courses

Fundamentals of Multimodal ML: A Taxonomy & Open Challenges

Paul Liang, Amir Zadeh and Louis-Philippe Morency

- ☑ 6 core challenges
- ☑ 50+ taxonomic classes
- ☑ 600+ referenced papers

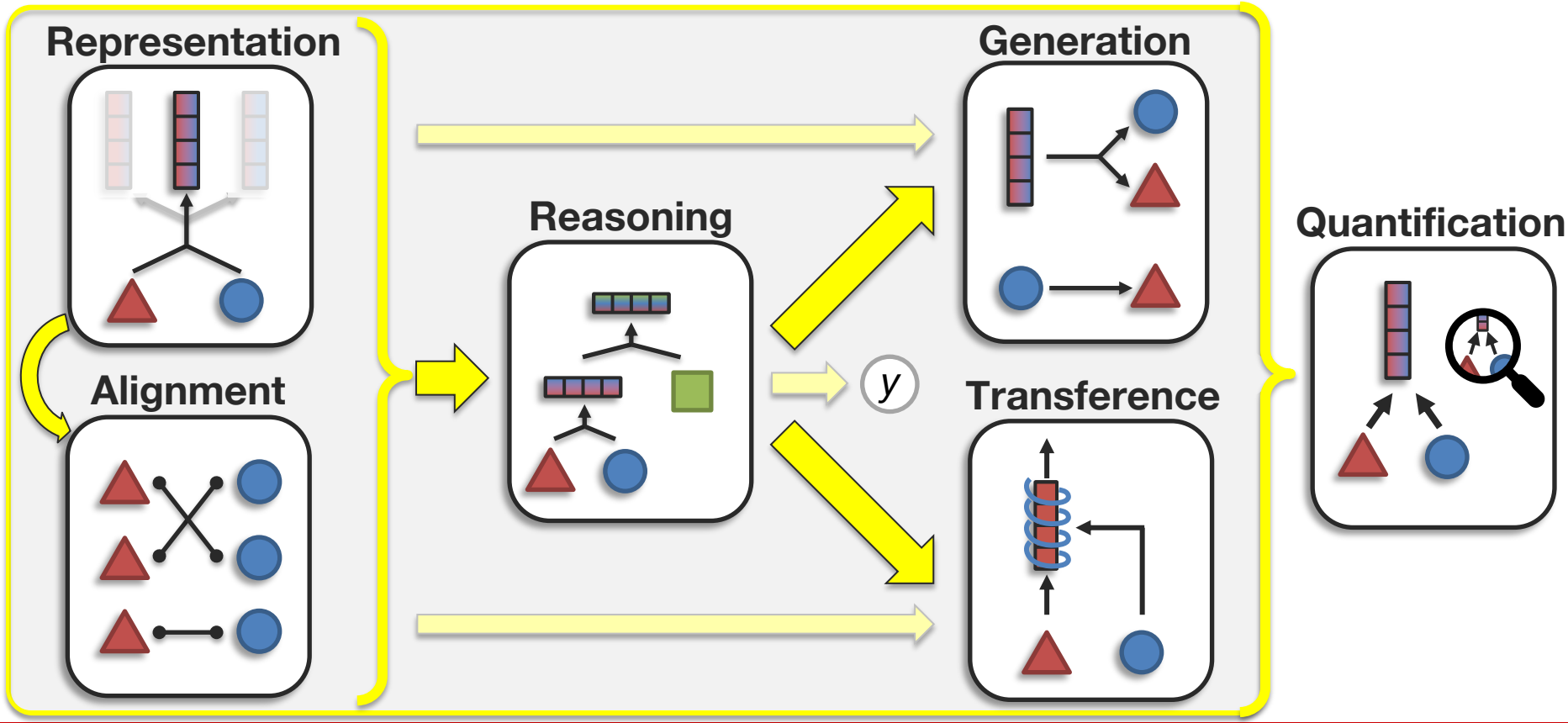
Tutorials: ICML 2023, CVPR 2022, NAACL 2022…

Graduate-level courses:

Multimodal Machine learning (11th edition)
https://cmu-multicomp-lab.github.io/mmml-course/fall2023/

Advanced Topics in Multimodal ML
https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

Core Multimodal Challenges



Course Syllabus

Learning Objectives

1. Study recent technical achievements in multimodal research
2. Improve critical and creative thinking skills
3. Understand future research challenges in multimodal
4. Explore new research ideas in multimodal learning

Two Versions: 6-credits and 12-credits

- 6-credit version:
  - Reading assignments
  - Small group discussions
  - Synopsis leads
- 12-credit version
  - Same 6-credit expectations + a high-quality research project:
    - Proposal with literature review
    - Midterm and final reports
    - Bi-weekly updates

Course Topics *(subject to change, based on student interests and course discussions)*

Week 1 (1/16): Introduction

Week 2 (1/23): Foundations 1: Dimensions of heterogeneity

Week 3 (1/30): Foundations 2: Modality connections

Week 4 (2/6): Foundations 3: Modality interactions

Week 5 (2/13): Multimodal LLMs 1: Data, pretraining, scaling

Week 6 (2/20): Multimodal LLMs 2: Fine-tuning, aligning, merging

Week 7 (2/27): Multimodal LLMs 3: Generative models

Week 8 (3/5): *No classes – Spring break*

Course Topics *(subject to change, based on student interests and course discussions)*

Week 9 (3/12): Interaction 1: Neuro-symbolic reasoning

Week 10 (3/19): Interaction 2: Embodied AI & planning

Week 11 (3/26): Interaction 3: Pragmatics & human-in-the-loop

Week 12 (4/2): Ethics and safety

Week 13 (4/9): Efficiency

Week 14 (4/16): Open discussion

Week 15 (4/23): Report presentations

Reading Assignments

- 12 readings assignments, with usually 2 required papers and some suggested (but optional) papers, and 5-6 discussion probes.
- Three main assignment parts (due Sunday night before discussion):
  - **Reading notes:** Read the assigned papers and summarize the main take-away points of each paper
    - Optional: if you have clarification questions about the papers
  - **Paper scouting:** Scout for extra papers, blog posts or other resources related to these question probes
  - **Discussion points:** Reflect on the question probes related to the reading papers and prepare discussion points.

How Each Weekly Tuesday Class will Happen

- Joint portion (about 15 mins)
  - Short presentation presenting the scouted papers and answering student questions about the required papers
- Separate discussion groups (about 1 hour)
  - Two groups of 8-10 students, one instructor per group
  - Round-table discussions: Discuss the research question probes. Each student is expected to actively participate in this discussion.
  - Two note-takers per discussion groups (alternating note-taking)

Discussion Roles

Reading leads (1 per discussion group, 2 total per week):

1. Short presentation (10-15 mins), done Sunday night - Tuesday
    a) Answer questions from other students
    b) Summarize and highlight scouted papers
2. Help with note-taking during discussions

Synopsis leads (1 per discussion group, 2 total per week):

1. Note-taking during discussions
2. Prepare discussion synopsis, done Tuesday - Monday
    a) Merge notes from both groups
    b) Summarize the main discussion points
    c) Organize into an overview schema, table or figure

Grading Scheme for 6-credit Version

- Reading assignments 40%
  - 6 points per reading assignment session
    - **1 point** for scouting relevant resources
    - **2 points** for take-away messages from assigned papers
    - **3 points** for reflections and thoughts on open discussion probes
  - Top 10 scores kept for final grade (total 12 assignments)

Grading Scheme for 6-credit Version

- Participation and discussions 40%
  - 4 points per discussion session
    - **2 points** for the insight and quality of the shared discussion points
    - **2 points** for interactivity and participation as follow-up to other's questions and suggestions.
  - Top 10 scores kept for final grade (total 12 discussions)

Grading Scheme for 6-credit Version

- Reading and synopsis leads 20%
  - Reading leads: 5 points for each presentation
    - **4 points** for creating and presenting the short presentation at the beginning of the course addressing the clarification points.
    - **1 points** for helping take notes of observations and points made during small group discussions
  - Synopsis leads: 5 points for each synopsis
    - **1 points** for taking notes during small group discussions. These notes should be posted on Piazza for all students.
    - **4 points** for creating the synopsis to summarize the main take-home messages of these discussions
  - Top 2 scores are kept for final grade

A Typical Week

- Previous Wednesday - @All reading assignment released
- Previous Sunday - @All reading assignment due
- Monday - @Reading leads make slides for clarifications + scouted papers
- **Tuesday** - @Reading leads present slides
- **Tuesday** - @All discussion in 2 groups
- **Tuesday** - @Synopsis leads take notes with help from @Reading leads
- **Tuesday** - @Reading leads submit slides for grading
- Thursday - @Synopsis leads submit 2 sets of notes
- Next Monday - @Synopsis leads merge notes and create coherent synopsis

What weeks would you prefer to lead reading & synopsis?

Week 2 (1/23): **Foundations 1: Heterogeneity**
Week 3 (1/30): **Foundations 2: Connections**
Week 4 (2/6): **Foundations 3: Interactions**
Week 5 (2/13): **Multimodal LLMs 1: Data, pretraining, scaling**
Week 6 (2/20): **Multimodal LLMs 2: Fine-tuning, aligning, merging**
Week 7 (2/27): **Multimodal LLMs 3: Generative models**
Week 9 (3/12): **Interaction 1: Neuro-symbolic reasoning**
Week 10 (3/19): **Interaction 2: Embodied AI and planning**
Week 11 (3/26): **Interaction 3: Pragmatics and human-in-the-loop**
Week 12 (4/2): **Ethics and safety**
Week 13 (4/9): **Efficiency**
Week 14 (4/16): **Open challenges**

Research Course Project (12-credit version)

Similar in spirit to a 6-credit independent study project

Project teams of 2 or 3 students

Final report should be like a research paper

Expected to explore new research ideas

Regular meetings with instructors on Thursday

Foundations of Heterogeneity

Motivation: Key concept in multimodal, but its definition and implications on modeling and training are not well understood

Challenges:
- What are different types and how to measure heterogeneity?
- How do different types of heterogeneity affect training? Model design? Evaluation?
- Modality-general model with modality-specific components that are automatically activated depending on heterogeneity?
- Modality tradeoffs & dynamic modality selection; benefits and risks of modalities
- Explain multimodal phenomenon: optimization challenges, modality collapse, modality overfitting

Potential models and dataset to start with
- Lots of literature on distribution shifts, federated learning, transfer learning, can they help?
- HighMMT, Gato, MultiBench etc resources with many many modalities

Foundations of Connections

Motivation: Key concept in multimodal, but its definition and implications on modeling and training are not well understood, esp fine-grained, compositionality, beyond co-occurrence

Challenges:
- Fine-grained connections between image and text, compositionality of connections, noisy connections. How to do representation learning and pre-training?
- Connections which are not due to exactly same meaning, e.g., various relationships, causal connections
- New ways to do contrastive and non-contrastive learning, connections with mutual info and connections.

Potential models and dataset to start with
- Traditional work on image-text relationships, recent compositional benchmarks eg. winoground
- Literature on causal relationship discovery
- Non-contrastive representation learning papers -> connection with mutual information?

Foundations of Interactions

Motivation: Key concept in multimodal, but its definition and implications on modeling and training are not well understood

Challenges:
- Formal measures of interactions and quantifying interactions in datasets and learned by models.
- How can we prove which models are most suitable for what types of interactions, how to do model selection.
- New models for new types of interactions eg., incorporating tensors and higher-order interactions into SOTA transformer models

Potential models and dataset to start with
- EMAP paper, literature on partial information decomposition, visualizing feature interactions
- How to generate synthetic data of various interactions. Bit-wise datasets

Understanding multimodal LLM uncertainty + human-in-the-loop

Motivation: Large multimodal models are black-box, we don't know what they know or don't know, when they're lying or hallucinating etc…

Challenges:
- Modeling and conveying model uncertainty – text input uncertainty, visual uncertainty, multimodal uncertainty? cross-modal interaction uncertainty?
- Detecting what models know or don't know, when they might be hallucinating.
- Asking for human clarification, human-in-the-loop, types of human feedback and ways to learn from human feedback. Multimodal – visual feedback, text feedback, cross-modal interaction feedback, etc.

Potential models and dataset to start with
- MMHal-Bench: https://arxiv.org/pdf/2309.14525.pdf aligning multimodal LLMs
- HACL: https://arxiv.org/pdf/2312.06968.pdf hallucination + LLM

Multimodal + Social

Motivation: Many social behaviors are multimodal in nature, socially intelligent AI requires social inference, social commonsense, social interaction.

Challenges:
- Social interaction between human beings are complicated to define and too diverse.
- Social concepts like deception and cheating are hard to detect even for human beings
- Social scenes are noisy and requires long-range multimodal understanding
- Complicated social understanding requires Theory-of-Mind ability to perform reasoning

Potential models and dataset to start with
- Multimodal WereWolf: https://persuasion-deductiongame.socialai-data.org/
- Ego4D: https://arxiv.org/abs/2110.07058
- MMToM-QA: https://openreview.net/pdf?id=jbLM1yvxaL
- 11866 Artificial Social Intelligence: https://cmu-multicomp-lab.github.io/asi-course/spring2023/

Interactive multimodal web agents

Motivation: Getting multimodal models grounded in the web or other virtual worlds, help humans with computer tasks.

Challenges:
- Web visual understanding is quite different from natural image understanding
- Instructions and language grounded in web images, tools, APIs
- Asking for human clarification, human-in-the-loop
- Search over environment and planning

Potential models and dataset to start with
- WebArena: https://arxiv.org/pdf/2307.13854.pdf
- AgentBench: https://arxiv.org/pdf/2308.03688.pdf
- ToolFormer: https://arxiv.org/abs/2302.04761
- SeeAct: https://osu-nlp-group.github.io/SeeAct/

Multimodal + Embodiment

Motivation: Embodiment requires perception, reasoning, and interaction - need to understand the influence of its actions on the world (i.e. long-term states, rewards)

Challenges:
- Multimodal + reinforcement learning
- How to learn models of the world with physical commonsense
- Simulator in embodied environment has too much freedom and hard to define the correct action space

Potential models and dataset to start with
- Virtual Home: http://virtual-home.org/paper/virtualhome.pdf
- Habitat 3.0 https://ai.meta.com/static-resource/habitat3
- RoboThor: https://ai2thor.allenai.org/robothor
- LangSuite-E: https://github.com/bigai-nlco/langsuite
- Language models and world models: https://arxiv.org/pdf/2305.10626.pdf
- Minecraft (MineDojo) and Voyager model: https://voyager.minedojo.org/

Improving multimodal LLM reasoning

Motivation: Robust, reliable, interpretable reasoning in multimodal LLMs.

Challenges:
- Fine-grained and compositional reasoning
- Neuro-symbolic reasoning

Potential models and dataset to start with

- Can LLMs actually reason and plan?
- Code for VQA: CodeVQA: https://arxiv.org/pdf/2306.05392.pdf, VisProg: https://prior.allenai.org/projects/visprog, Viper: https://viper.cs.columbia.edu/
- Cola: https://openreview.net/pdf?id=kdHpWogtX6Y
- NLVR2: https://arxiv.org/abs/1811.00491
- Reference games: https://mcgill-nlp.github.io/imagecode/, https://github.com/Alab-NII/onecommon, https://dmg-photobook.github.io/

Ethics and safety

Motivation: Large multimodal models are can emit unsafe text content, generate or retrieve biased images

Challenges:
- Taxonomizing types of biases: text, vision, cross-modal interaction, generation, etc.
- Tracing biases to pretraining data, seeing how bias can be amplified during training, fine-tuning, etc.
- Various ways of mitigating biases

Potential models and dataset to start with
- Many works on fairness in LLMs -> how to extend to multimodal?
- Mitigating bias in text generation, image-captioning, image generation

New Modalities

Motivation: Many tasks of real-world impact go beyond image and text

Challenges:
- Multimodal with non-deep-learning effective modalities (e.g., tabular, time-series)
- Multimodal deep learning + time-series analysis models or + tabular models
- Multimodal with low-resource modalities
- Generating faces, gestures, virtual humans

Potential models and dataset to start with
- Brain EEG Signal: https://arxiv.org/abs/2306.16934
- Speech: https://arxiv.org/pdf/2310.02050.pdf
- Facial Motion: https://arxiv.org/abs/2308.10897
- Tactile: https://arxiv.org/pdf/2204.00117.pdf

Bi-weekly Project Meetings and Updates

- Required meetings on a bi-weekly basis
  - About 20 minutes per meeting during Thursday class time
  - Primary mentor (Paul or Daniel) for each team
- Bi-weekly written updates
  - Either Google Slides (preferred) or Google Docs
  - Due Tuesdays at 9pm before the meeting
  - Some expectations for each bi-weekly update (see next slide)
- Alternate weeks: optional meetings with either mentor
  - No written update required, but suggested

Schedule for Bi-Weekly Written Updates and Reports

- Week 3: Pre-proposal details with literature review
- Week 5: **Proposal report:** baseline results and new ideas
- Week 7: Initial implementation of new ideas
- *Week 8: Spring break (no meetings, no work, relax* 🙂*)*
- Week 10: **Midterm report:** first complete round of results for idea
- Week 12: Updated results for research idea
- Week 14: Error analysis, ablations, and visualizations
- Week 15: **Project presentations**
- Week 16: **Final report**

Course Project Timeline

- **Project preferences** (Due Friday 1/19 at 9pm ET) – Share your interests about research projects, to help with team matching.
- **Pre-proposal** (Due Tuesday 1/30 at 9pm ET) – You should have selected your teammates, have ideas about your dataset and task.
- **Proposal and Literature Review** (Due Tuesday 2/13 at 9pm ET) – Research ideas, review of relevant papers and initial results
- **Midterm report** (Due Tuesday 3/19 at 9pm ET) – Intermediate report documenting the updated results exploring your research ideas.
- **Final report** (Due Tuesday 4/30 at 9pm ET) – Final report describing explored research ideas, with results, analysis and discussion.

Grading Scheme for 12-credit Version

- Grading breakdown of the 6-unit version will be scaled to 50%.

- The second 50% comes from the course project:
  - Proposal report 10%
  - Midterm report 20%
  - Final report 25%
  - Final presentation 15%
  - Bi-weekly written updates 30%
    - 10 points per update, top 3 scores kept for final grade (out of 4 updates)

Absences and Late Submissions

- Lectures are not recorded, students expected to attend live
  - If you plan to miss more than one lecture this semester, let us know as soon as possible.
- Reading assignment wildcards (3 per students)
  - 24-hours extension, max 1 per week
- Project assignment wildcards (2 per teams)
  - 24-hours extension, can be used together

Course Websites

- Piazza
  - For course announcements and assignments

  https://piazza.com/cmu/spring2024/11877/info
- CMU Canvas
  - For assignment submissions and grading

  https://canvas.cmu.edu/courses/39063
- Course website
  - A public version of the course information
    - Discussion synopsis will be posted here

  https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2024/

Assignments for This Coming Week

Week 2 reading assignment (Due Sunday 1/21 at 9pm ET)

- Detailed instructions will be posted on Piazza
  - Required paper: Geometric deep learning, a unified paradigm to reason about structure, invariance, properties, and inductive biases in each modality.
  - Suggested papers: Useful dimensions of heterogeneity in domain adaptation, transfer learning, multitask learning, quantifying dimensions of heterogeneity.

For students taking the 12-credit version:

- Project preference form (Due Friday 1/19 at 9pm ET)
  - To help with team matching
  - Google Form link will be available on Piazza